Using time series analysis to analyze trends in concentration

Skip Vecchia

U.S. Geological Survey

Bismarck, ND

What's the difference between regression and time series analysis?

In a regression model, the primary goal is to estimate the "expected value" of the independent variable

Regression model: $Y = E[Y|X] + \varepsilon$

- Want to estimate E[Y|X] (the conditional expectation of Y given X)
- The errors generally are treated as "noise", whose statistical properties are used to evaluate uncertainty in E[Y|X]
- The errors may be uncorrelated (ordinary least-squares) or they may have some "nuisance" correlation that is accounted for in fitting the regression model (generalized least squares)

What's the difference between regression and time series analysis? (continued)

In time series analysis, the primary objective is to predict the independent variable

Time series model:

$$\mathbf{Y(t) = E[Y(t)|\ \{X(t-k\delta),\ Y(t-(k+1)\delta),\ k = 0, 1, \ldots\}] + \varepsilon(t\ ;\ \delta)}$$

- Y(t) is the variable to be predicted (eg, log of sediment conc.)
- X(t) is an "explanatory" time series (eg, log of discharge)
- δ is the time increment (eg, one day, week, or month)
- $\varepsilon(t\ ;\ \delta)$ is the prediction error for time t (note: $\varepsilon(t\ ;\ \delta)$ decreases to zero as δ gets small)

QWTREND: Detecting trends in water-quality data using time series analysis

- Computer program for analyzing trends in concentration
- Based on a parametric time series model for concentration and streamflow, described below
- Software packages (both stand-alone and S-Plus version) should be available for public distribution in Fall 2004

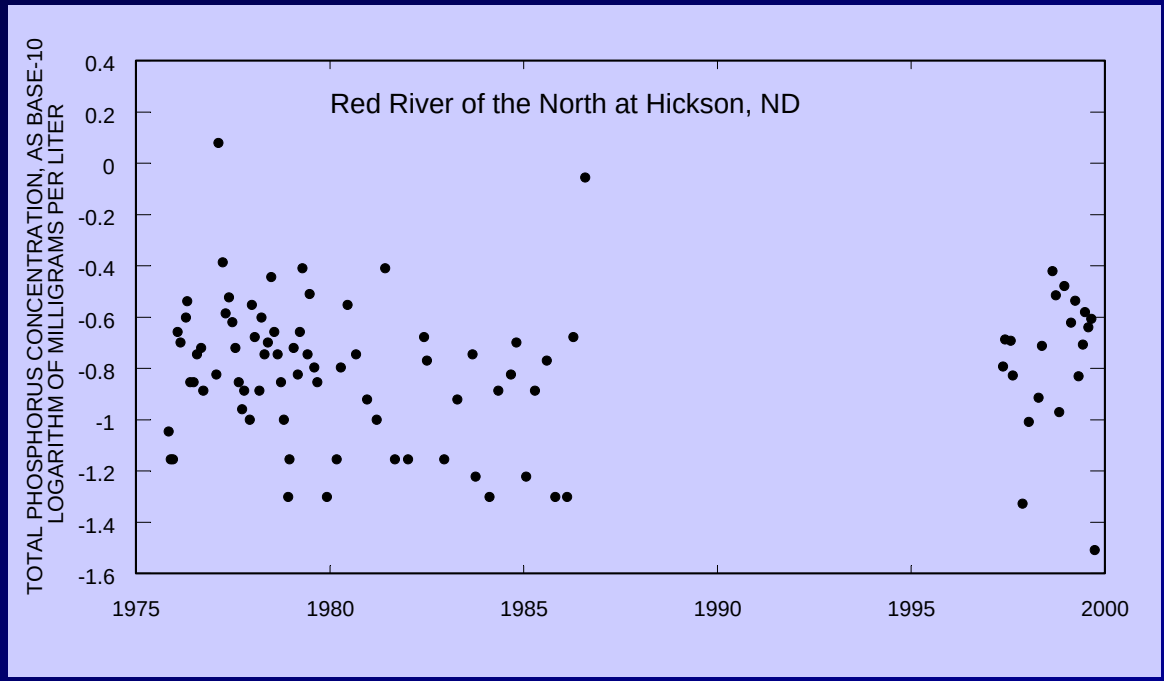
Supplemental reading for QWTREND

1. Vecchia, 2003, Water-Quality Trend Analysis and Sampling Design for Streams in North Dakota: USGS WRIR 03-4094 (available online at http://nd.water.usgs.gov)
2. Trench and Vecchia, 2002, Water-Quality Trend Analysis and Sampling Design for Streams in Connecticut: USGS WRIR 02-4011
3. Vecchia, 2003, Relation Between Climate Variability and Stream Water Quality in the Continental United States: Proceedings of the American Institute of Hydrology (preprint available by request from avecchia@usgs.gov)

Data Requirements for QWTREND

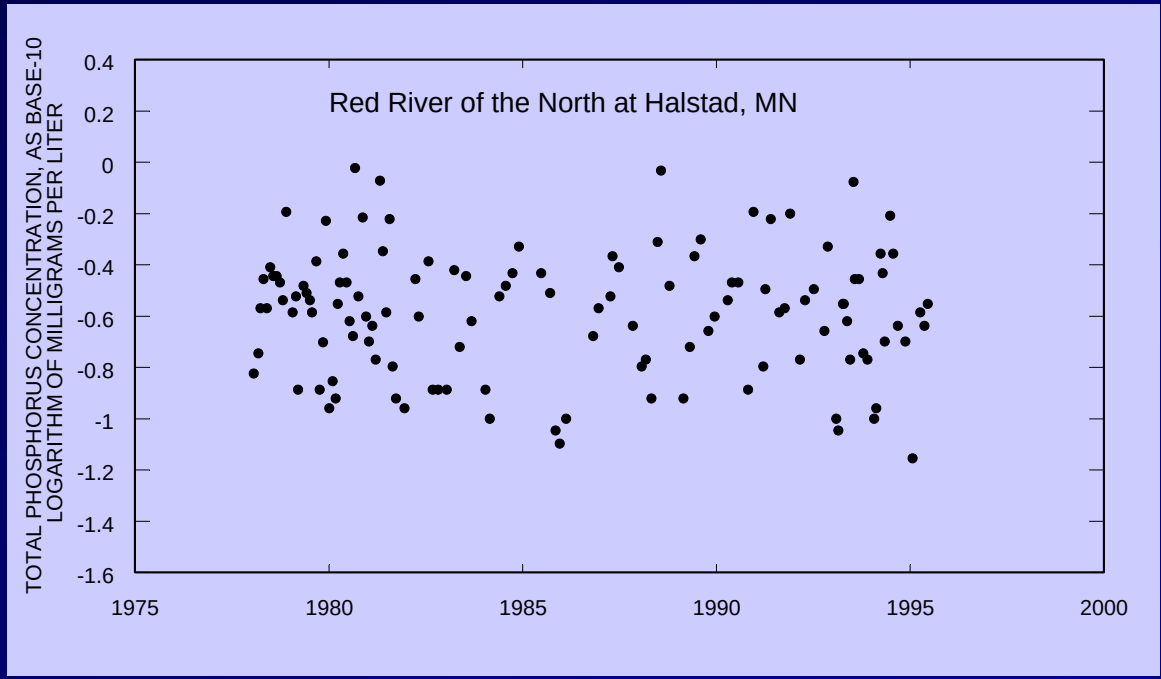
- Record length of at least 15 years
- Average of at least 4 samples per year (sampling frequency may vary from year-to-year)
- For each 3-month season (Jan-Mar, Feb-Apr, Mar-May, …, Dec-Feb), at least 1 sample in each of 10 separate years
- No more than 10 percent of values below detection limit
- Full record of daily streamflow from 5 years before the first water quality sample through the end of the record

Data Requirements for QWTREND (2)



These data failed requirement 3 (too many observations clustered at the ends)

Data Requirements for QWTREND (3)



These data satisfy the requirements

General time series model for QWTREND

Log C(t) = M + SEAS + SRA +TND + PRDEV + NOISE

- M is the long term mean (constant)
- SEAS is seasonality (periodic function with period 1 year)
- SRA is the streamflow-related anomaly
- TND is the trend
- PRDEV is a predicted deviation
- NOISE is the prediction error

Streamflow-related anomaly

SRA = E [log C(t) – M – SEAS – TND | X(t), X(t-1), …]

- X(t) = log Q(t) – E [log Q(t)]
- Q(t) is daily streamflow

Best (minimum-variance, unbiased) predictor of the deviation of log-transformed concentration from "basic conditions" (M + SEAS + TND), based on past and present streamflow

This looks a lot like the standard flow-adjustment model we looked at earlier, except in the standard model we used only concurrent streamflow, X(t)

Trend

$TND = \beta_1 F_1(t) + \beta_2 F_2(t) + \ldots + \beta_k F_k(t)$

$F_1(t)$, $F_2(t)$, … , $F_k(t)$ are specified trend functions, for example:

- Step trends
- Piecewise-linear trends
- Splines
- Functions of known covariates, such as fertilizer application, crop production, population, etc.

Predicted deviation

$Y(t) = \log C(t) - M - SEAS - SRA - TND$ "deviation"

PRDEV = E [Y(t) | Non-missing Y(t-1), Y(t-2), …]

- Minimum variance, unbiased predictor of Y(t)
- A Kalman-filtering algorithm is used to compute PRDEV
- Computation is "tedious" because of all the missing values and because of the potentially complex serial correlation structure of the deviations

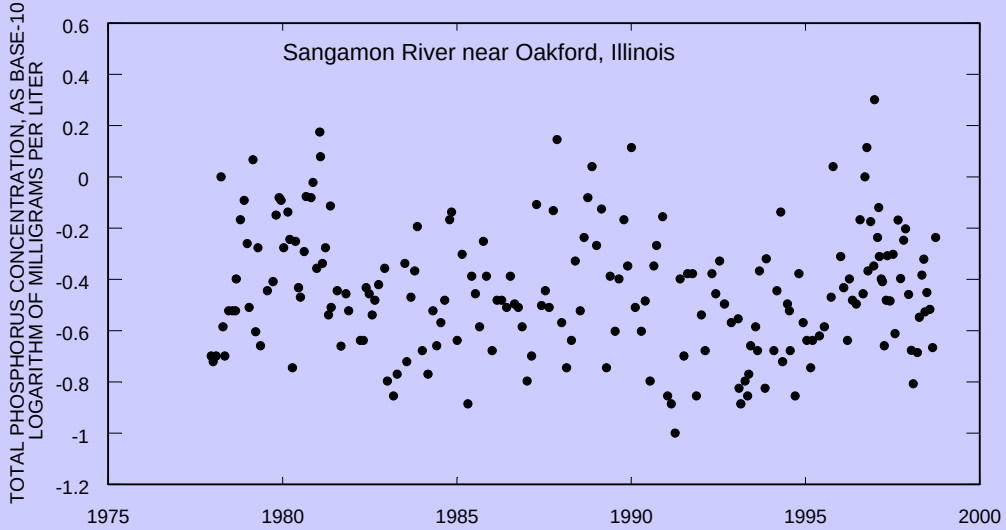
Filtered concentrations

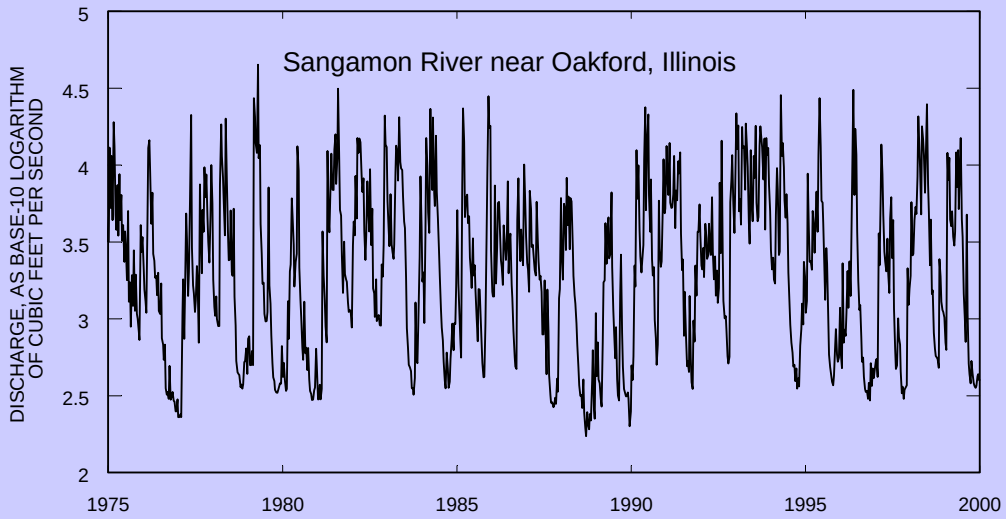
FILC(t) = log C(t) – M – SEAS – SRA – PRDEV

FILC(t) = TND + NOISE

- Similar to flow-adjusted concentrations, but less variability
- The time series model usually removes more variability than simple flow-adjustment, making trends easier to detect
- All parameters, including trend parameters and time series parameters, are estimated via maximum likelihood (allows trends to be analyzed using likelihood-ratio tests)

Example – Total Phosphorus

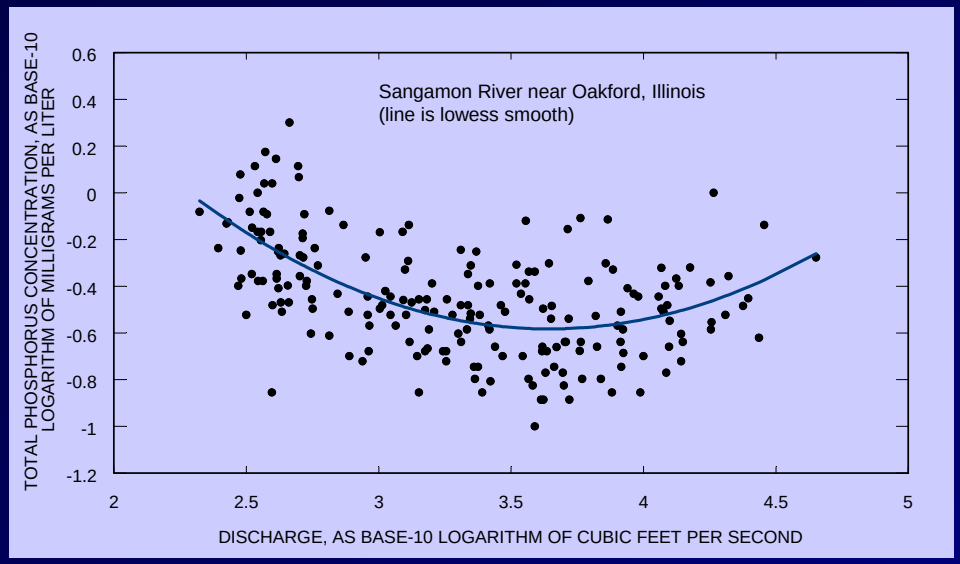




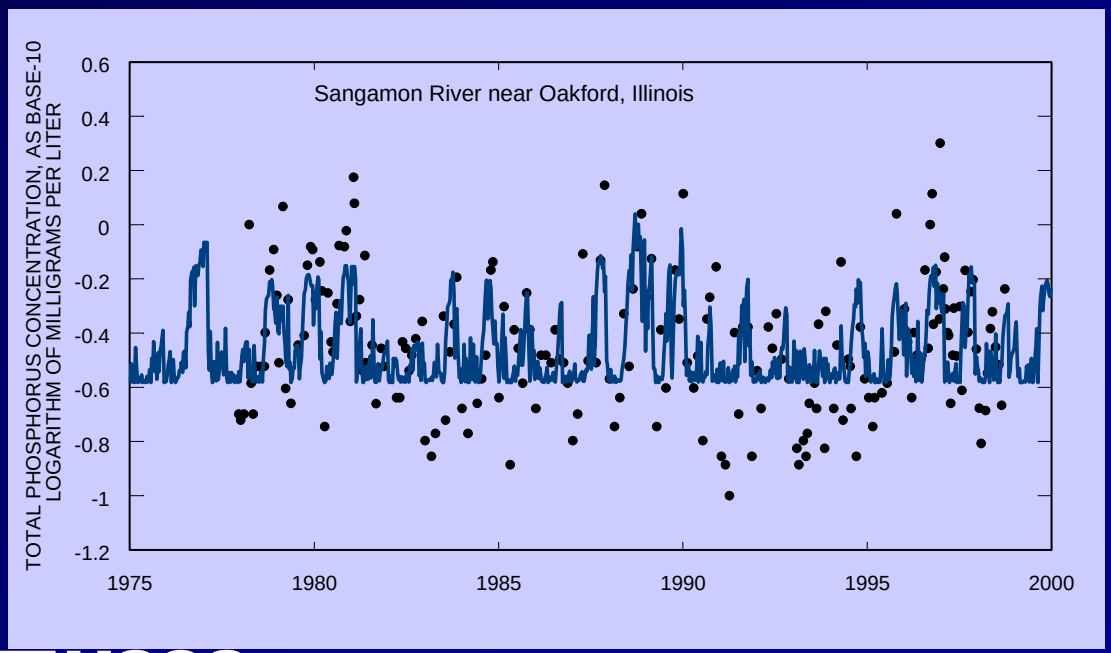
This is the same data we looked at previously (but log transformed)

This is the daily discharge data for this station

Example – Total Phosphorus (2)

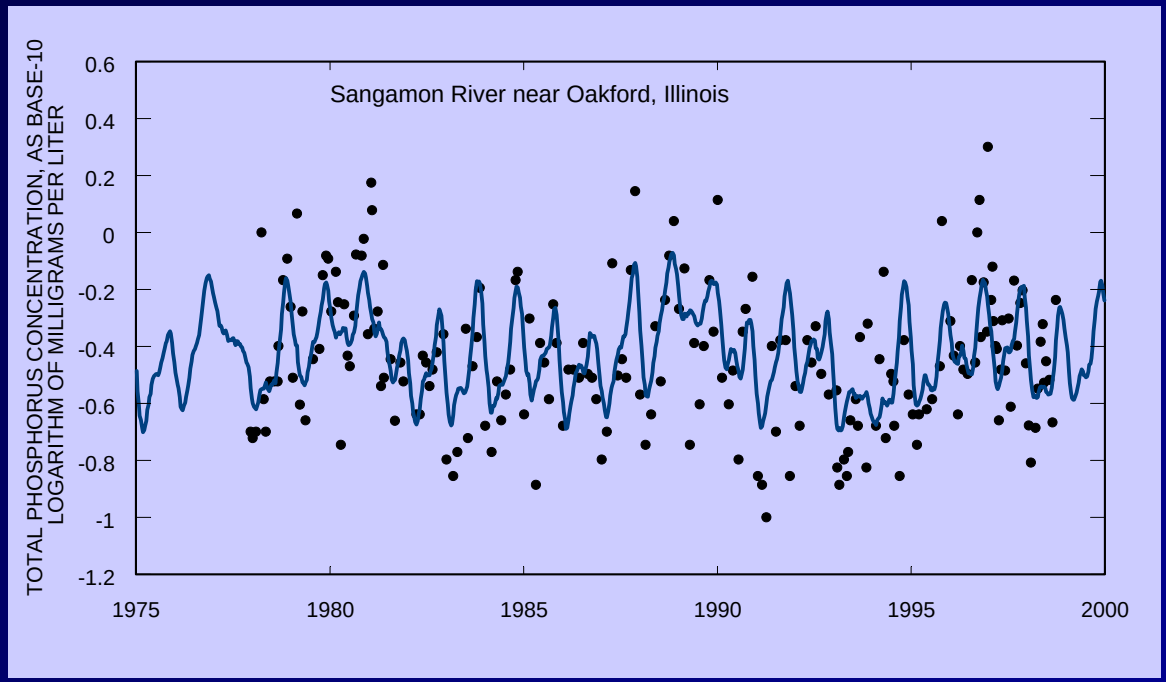


This shows the fitted regression line for log of conc. versus log of discharge (simple flow-adjustment model



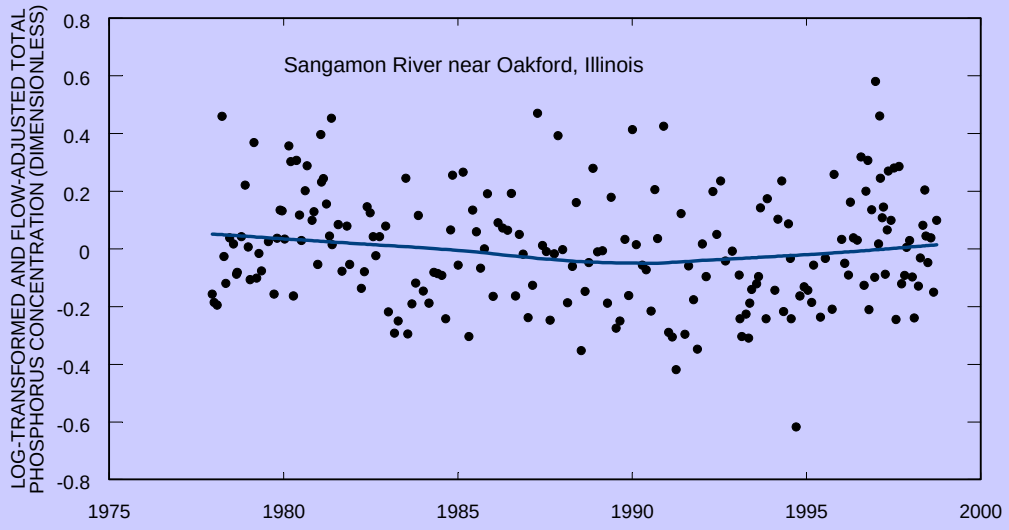
This line shows the time series of fitted values for concentration (one for each day)

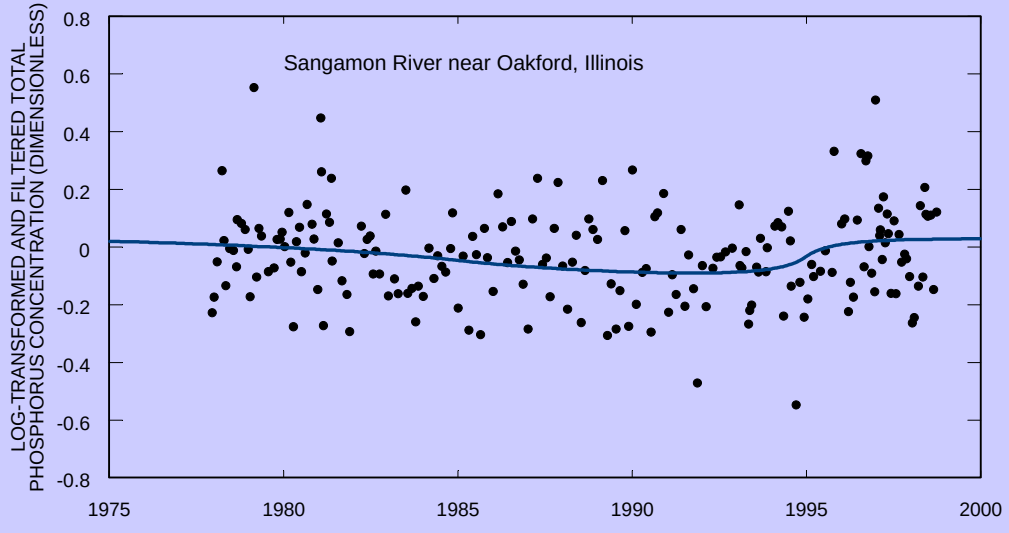
Example – Total Phosphorus (3)



This line shows the fitted streamflow-related anomaly (M + SRA) from QWTREND – note the difference from the previous slide.

Example – Total Phosphorus (4)

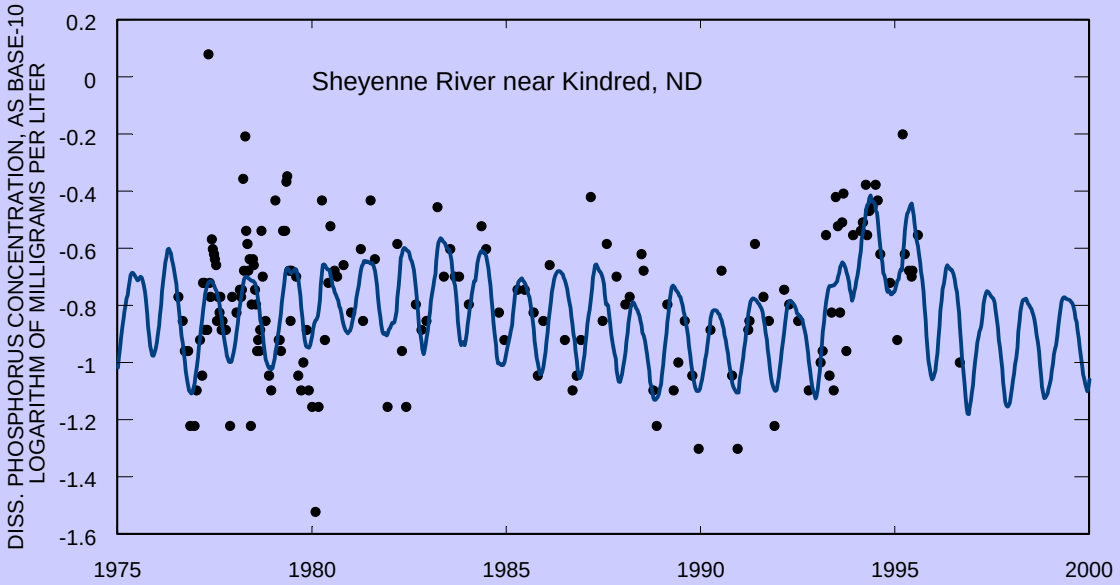


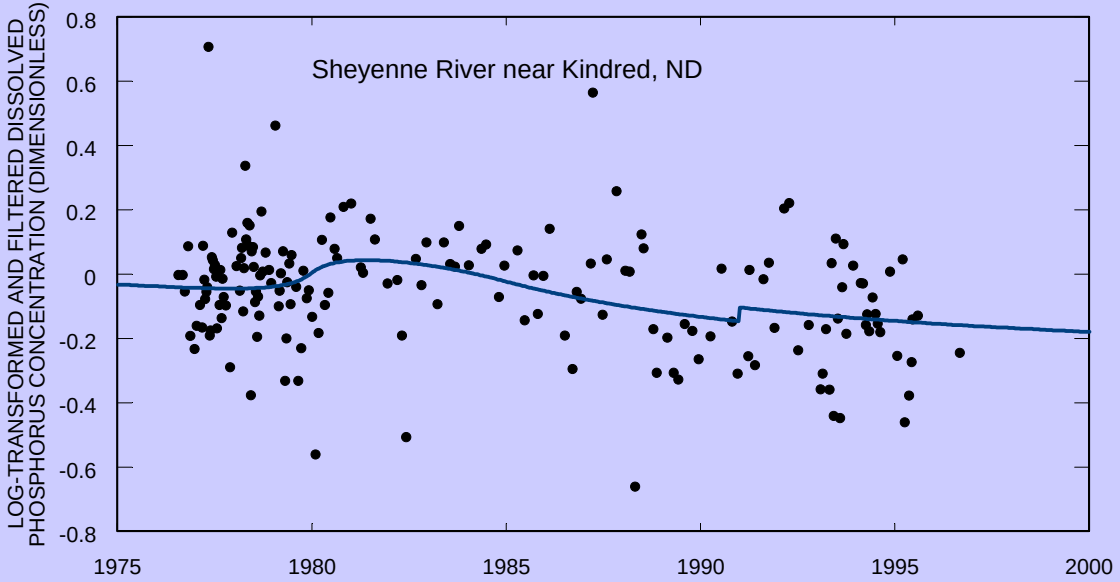


This is a plot of the flow-adjusted concentrations along with a lowess smooth. No significant trend was found using SEAKEN

This is a plot of the filtered concentrations (FILC) from QWTREND

The fitted trend is highly significant ($p<0.001$), using a likelihood ratio test





Another example

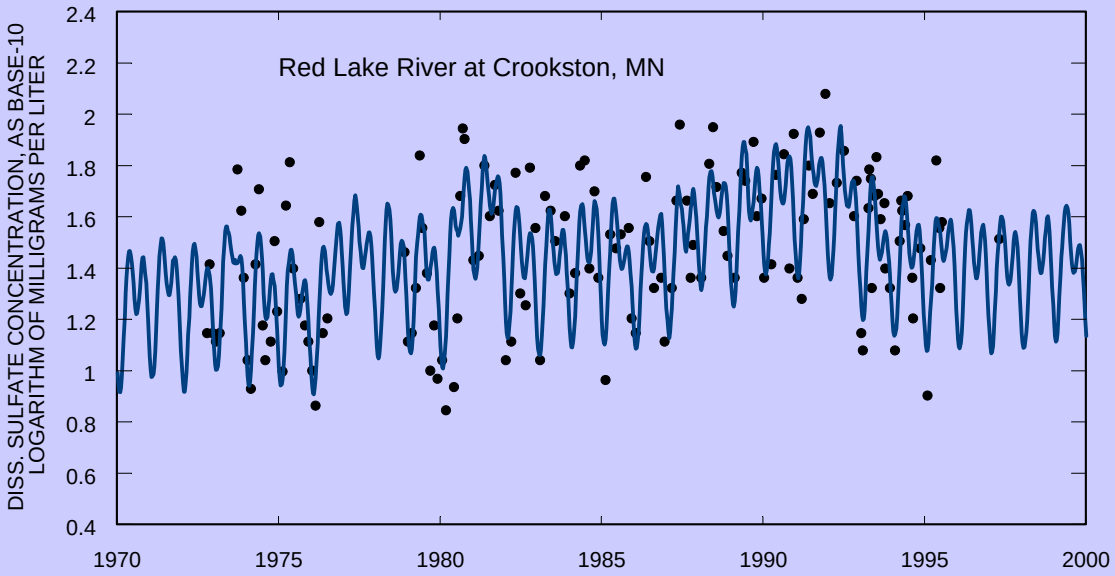
Points: dissolved phosphorus concentrations

Line: Fitted SRA plus trend

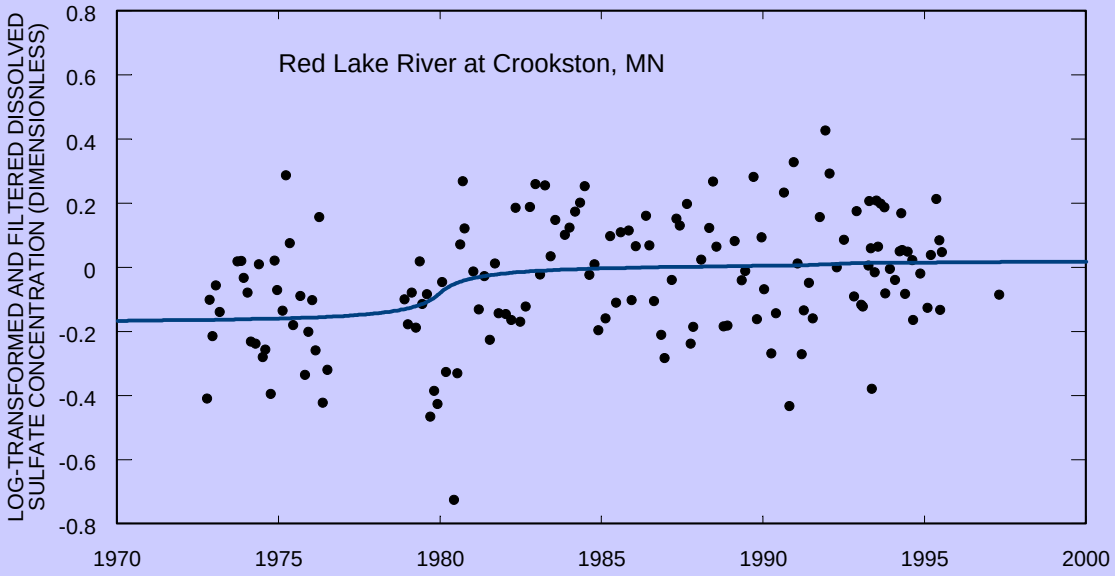
Points: filtered concentrations

Line: Fitted trend ($p<0.001$)

One more example



Dissolved sulfate concentrations (points) and fitted SRA plus trend (line)



Filtered concentrations (points) and fitted trend ($p<0.001$)